

Exploring the efficacy of molecular fragments of different complexity in computational SAR modeling

Albrecht Zimmermann, Björn Bringmann, Luc De Raedt

Abstract—An important first step in computational SAR modeling is to transform the compounds into a representation that can be processed by predictive modeling techniques. This is typically a feature vector where each feature indicates the presence or absence of a molecular fragment. While the traditional approach to SAR modeling employed size restricted fingerprints derived from path fragments, much research in recent years focussed on mining more complex graph based fragments. Today, there seems to be a growing consensus in the data mining community that these more expressive fragments should be more useful.

We question this consensus and show experimentally that fragments of low complexity, i.e. sequences, perform better than equally large sets of more complex ones, an effect we explain by pairwise correlation among fragments and the ability of a fragment set to encode compounds from different classes distinctly. The size restriction on these sets is based on ordering the fragments by class-correlation scores. In addition, we also evaluate the effects of using a significance value instead of a length restriction for path fragments and find a significant reduction in the number of features with little loss in performance.

I. INTRODUCTION

Structure-activity relationship (SAR) prediction is an important task in computational biochemistry. The aim is to predict the effect of compounds based on their structural characteristics – the second-order representation comprising the topological arrangement of atoms and bonds of the molecule.

For algorithms to process molecular data and build models to predict their activity, molecules have to be simplified by transforming them into a different representation. To this end, molecules are abstracted as *graphs* – networks of atoms linked to each other. A common approach to SAR consists of constructing fragments from individual or pairs of molecules, and subjecting those molecules to *fingerprinting* to gain a final representation that is more easily accessible for prediction algorithms such as *Support Vector Machines*. The *graph mining community*, on the other hand [1], approaches the construction of fragments slightly differently and while the differences are subtle, they can have significant effects.

In this paper, we build on earlier work [2] that aimed at generalizing the existing fingerprinting approach and explore how to derive the molecular fragments on which to base generalized fingerprints in a predictive setting. Specifically, we compare fragments of different complexity in terms of their usefulness. Additionally, we compare the use of fragments

selected according to their correlation with the target value to ones selected using a threshold on their length.

The paper is structured as follows: we first discuss the concept of fingerprints and its extension to *generalized fingerprints*, as well as differences in the complexity of fragments on which to base them. Following this, we lay out our methodology for the experimental comparisons in terms of complexity and selection criterion, on which we report afterwards. Finally, we discuss the observed phenomena and draw conclusions.

II. (GENERALIZED) FINGERPRINTS

The usual approach in computational biology/chemistry when using the second-order representation for SAR involves assessing the structural similarity of molecules. They are decomposed into sets of (potentially overlapping) fragments and the similarity of any two molecules evaluated comparing their respective fragments using *kernel functions* [3].

A variety of different fragments has been used in the literature, from paths/walks [4], [2], [5], via fragments with branches (trees) [6], to those with cycles (graphs) [7], [8], [9], [10]. Often, a new kernel function is proposed as well. These approaches share two characteristics: 1) they start from vertices (atoms) of individual or pairs of molecules, enumerating the paths starting from this vertex, or the neighborhood graphs surrounding it. 2) the fragments are size restricted, length restricted for paths (such as $0 \leq l \leq 8$ or $3 \leq l \leq 10$), or diameter restricted for graphs.

The resulting fragments are often used to map molecules to bit-vectors of a given size k (such as 512 or 1024), in a process called *fingerprinting*, involving the generation of b random integers that are mapped using a modulo k reduction. While values such as k , l , and b are based on empirical knowledge of biochemical practitioners, it has been shown that, e.g., different length restrictions can have a profound effect on the usefulness of derived fragments [11].

As an alternative to hashing, Swamidass *et al.* [2] proposed *generalized fingerprints (gfp)* in which the *explicit* size restriction on bit-strings is lifted. Thus, potential loss of information is avoided since *each* fragment is represented. Also, it becomes possible to use information that goes beyond presence, e.g. how often a fragment occurs in the data, which the authors exploited in proposing a kernel. However, hashing fragments to the same bit can weed out redundancy and such a representation potentially avoids the curse of dimensionality, and reduces memory requirements for the modeling step. The

Albrecht Zimmermann, Luc De Raedt, KU Leuven, first-name.lastname@cs.kuleuven.be

B. Bringmann, Deloitte & Touche GmbH, bbringmann@deloitte.de

retention of information seems to outweigh the benefits of hashing, since Swamidass *et al.* showed that *gfp* outperform *fp*, especially for smaller *fps*.

Their approach used path fragments, and in their conclusion they suggested the use of shallow trees as fragments from which to construct *gfp*. This coincides with trends in the data mining community where *graph mining* is touted as the tool of choice to derive fragments for SAR prediction. In contrast to this there have been claims that simpler features may well suffice [12], [13]. This assumption has been supported by recent work [8] that evaluated the efficacy of structures of different complexity against one another and found little, if any, advantage in using more complex structures such as graphs. It has to be remarked, however, that the latter work still constructs *size restricted* fragments from individual molecules.

In contrast to this, we have found in the past that sequential fragments are *more* useful than more complex ones [14]. Yet we construct fragments differently: they are not size restricted but chosen based on how well they correlate with the target variable, measured by χ^2 , a correlation that is evaluated on the entire data. The size restriction on fingerprints can either be enforced explicitly by taking the k best-correlating fragments, or implicitly by requiring a minimum correlation score.

We reproduce our experiments on new data and perform additional analysis to answer the following questions:

- Q1. Are fragments with low complexity as useful as more complex fragments and if so, what are the underlying phenomena?

Restricting the number of fragments used gives us a controlled setting in which to evaluate the efficacy of fragments from different fragment classes. By analyzing the encoding of the data that can be derived from the mined patterns, and the relation among patterns themselves, we gain an intuition as to why simpler structures are the better choice when the number of patterns is limited in the mining process. It is not obvious whether these results will transfer to a size restricted setting but we can answer a related question, namely:

- Q2. Is mining fragments using a significance threshold as good as the size restricted approach to building *gfps*?

The size restricted approach is equivalent to considering the occurrences of *all* fragments adhering to those size restrictions and using those that occur at least once. This can lead to an explosion in the number of enumerated fragments, *even* using length restriction on the patterns, as we will show. Arbitrarily increasing this threshold, on the other hand, might exclude interesting fragments, and as mentioned above, the effect of changing the size restrictions is not always predictable [11].

Analogously, minimally correlating fragments can be considered to consist of at least two atoms and to adhere to a correlation constraint. Changing the size constraint can still have unpredictable results whereas changing the correlation threshold has a clear interpretation. Consequently, in a final experiment, we compare the effect of increasing the number of fragments of low complexity by lowering the mining threshold, showing the increasing quality of the encoding (and its consequences for the quality of the classification model), and contrasting their usefulness with length restricted fragments.

III. APPROACH

We use substructures that correlate with one of two target classes (e.g. *active* and *inactive*) – and therefore discriminate among the two. Techniques exist for mining top- k substructures according to convex measures such as χ^2 or *Information Gain* while still pruning large parts of the search space. Similar search strategies can be used to find all substructures with a score above a user defined threshold. Please note that in this work we only use χ^2 since earlier work showed that this leads to better results than employing Information Gain [15].

Regarding chemical compounds, there exist three very well studied types of substructures, namely:

- \mathcal{L}_G *subgraphs*, most expressive, but expensive to mine;
- \mathcal{L}_T *subtrees*, can represent anything but cycles;
- \mathcal{L}_S *subsequences*, least expressive, rather easy to mine.

The relation $\mathcal{L}_S \subset \mathcal{L}_T \subset \mathcal{L}_G$ holds, implying that $|\mathcal{L}_S| \leq |\mathcal{L}_T| \leq |\mathcal{L}_G|$. Note that sequences are slightly different from paths as used by Swamidass *et al.* as they only allow a bijective mapping of the nodes and edges from the fragment to the data, i.e. a vertex can occur at most once in a sequence. Our first question is concerned with comparing these three types of structures w.r.t. their value in terms of predictive accuracy. To carry out this task, we extract a number of substructures from the data, and use them to describe each of the seen or unseen chemical compounds. The molecules are transformed into generalized fingerprints indicating the substructures’ presence or absence. From the feature vectors a model for the activity of the compounds is learned.

Support vector machines (SVMs) have been used successfully for SAR problems and can filter out redundant/irrelevant features. We use the *Tanimoto kernel* that has been used to good effect on the NCI cancer data set we do our comparison on [2]. The data is encoded as undirected graphs, vertices labeled with their atom type, edges as single, double, or aromatic bonds. Hydrogen atoms are not encoded. The shortest possible sequence consists of a single edge, i.e. two atoms.

IV. EXPERIMENTAL EVALUATION

The NIC60 cancer data set is a popular data set for testing SAR predictions [2], [9]. It consists of approximately 4000 compounds that have been tested against 60 tumor cell lines. Each of the 60 subsets consists of around 3,500 compounds.

For each of the 60 cell lines comprising the NCI cancer data set, we performed a stratified 10-fold cross validation. Fragments are mined exclusively on the training folds since there is no information about class labels in the test data.

A. Comparing different complexity classes – *fps* of equal size

The classical approach to encoding molecules in *fps* lies in fixing the size k of the *fp* and hashing fragments to a bit-string of this length. In the technique we propose, mining significant fragments according to the χ^2 statistic, fixing the size of *fps* can be done by mining only the k highest-scoring fragments. In earlier work [14], we showed that for a fixed k , sequences were more effective in encoding data for classification purposes than trees or graphs. We repeat the

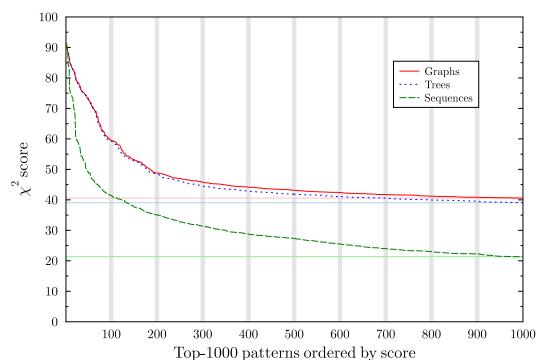


Fig. 4. Score distribution for the top-1000 fragments according to χ^2 of one representative training fold.

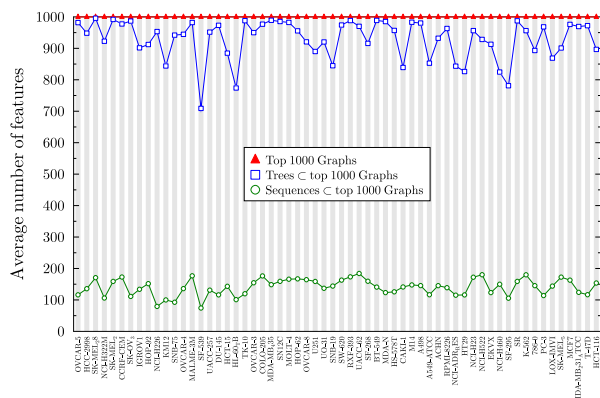


Fig. 5. Number of fragments remaining after the set is reduced using the 1000th-worst graph score.

To normalize the observed advantage of sequences with regard to diversity, we use the 1000th-best graph score (the red horizontal line) to crop the size of *fps* derived from tree and sequential fragments. We effectively obtain *generalized fingerprints*, without explicit size restriction of bit-vectors, similar to the ones used by Swamidas *et al.*, with the *length* restriction on fragments replaced with a minimum *significance* value. The number of fragments left is shown in Figure 5. This reduction is in fact rather severe, pushing the number of sequence fragments down to 10% – 20% of the original 1000.

As mentioned before, Figure 3 also shows the average number of sequential features per molecule in the top 1000 graphs, which is equivalent to using the reduced set of sequential features (bottom curve). In comparison to the second curve from the bottom, one can see that, again, the reduction is severe, yet not as severe as for the entire set of features, only dropping to 25% – 33% of the original number ($\sim 40 - 50$).

Reducing the number of features in this way leads to a very slight advantage for the graph-structured features w.r.t. AUC results (lower half of Figure 6). In fact, according to the Wilcoxon test, there are only two significant differences, even though so many fewer sequences than graphs are used.

Similarly to the results described in the preceding section, the decrease in accuracy goes along with an increase in the number of correspondences, as shown in the upper half of

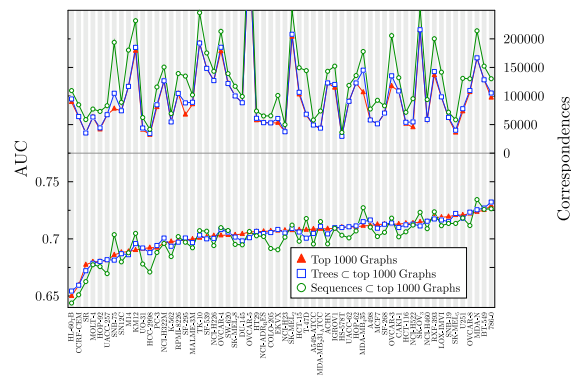


Fig. 6. AUC results of SVM-classification on encodings derived from the less complex fragments contained in the top-1000 graphs and number of correspondences in each encoding.

Figure 6. It is important to note the trade-off among the number of features and the quality of the feature set. The number of correspondences are rather similar for each data set for all structure types, despite the differences in total number of fragments. This indicates that the large sets of tree- and graph-structured fragments still show much redundancy, which in turn means that while additional complexity allows for some more diversity for a given threshold, the gain is relatively small compared to simply increasing the number of fragments.

C. Increasing *gfp*-size by lowering the mining threshold

Increasing the number of features improves the chance that molecules from different classes are encoded in a way that allows to distinguish those classes. Tree and graph mining being far more expensive than sequence mining [14], it is unrealistic to try and mine large amounts of complex fragments. Also, using all graphs which have a χ^2 -score exceeding a given threshold does little to increase diversity over sequences.

The computational complexity of fragment mining arises from the need to systematically explore a large search space of potentially interesting fragments and count their occurrences in the data. Approaches that start from individual molecules avoid this bottleneck, yet while the fragments derived in that manner can be used for assessing molecules' similarities this is often the extent of their usefulness, especially since they are often tied to their respective kernel functions. Fragments correlating with the target value, on the other hand, capture information about the relationship of structure and activity themselves and can be analyzed independently from the modeling step.

In a third experiment we thus mine sequences which have a χ^2 -score exceeding the (unadjusted) 95%, 99%, and 99.9% p-values, respectively. As expected, Figure 8 shows a direct relationship between lowering the significance threshold and the number of features mined. More features also leads to fewer correspondences which lead then to corresponding AUC values (Figure 7). According to the Wilcoxon Test, using the p-value for 99% improves significantly on the 99.9% value in 36 cases. Using the 95% value increases this to 45 (and improves in 17 cases on the 99% value).

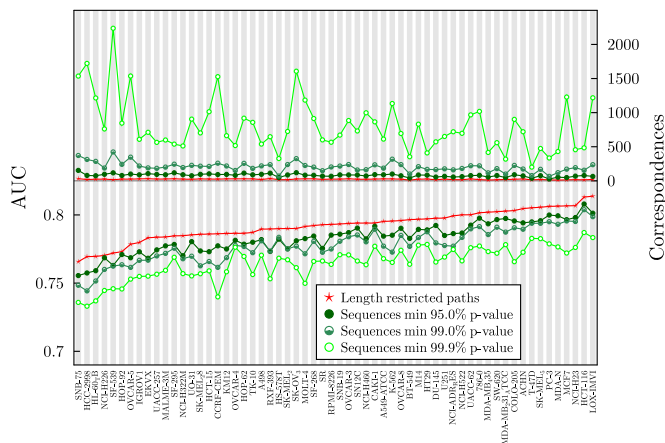


Fig. 7. AUC for sequences based on three different significance thresholds and length-restricted paths of frequency 1 along with the number of correspondences in each encoding.

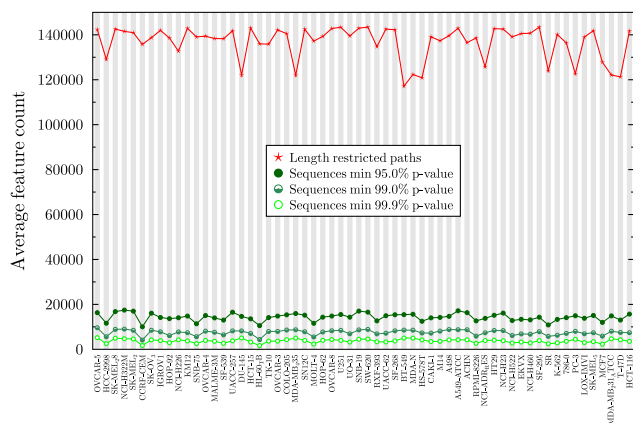


Fig. 8. Number of sequential-fragments for three different significance thresholds and length-restricted paths of frequency 1.

D. Comparing techniques for determining *gfp*-size: significance versus length-restriction

The preceding experiments had the main purpose of assessing the usefulness of graph-, tree- and sequential fragments for *gfps*, chosen by the significance of their χ^2 score. Swamidass *et al.* [2] use *gfps* whose number is determined by a length restriction – all fragments are paths of maximal length 10 occurring in at least *one* molecule in the training data. This approach gives rise to more than one hundred thousand features (the top graph in Figure 8), significantly more than result even from using the permissive 95% p-value. According to the results described above, this should allow those feature sets to encode all molecules distinctively. Indeed, the low frequency threshold means that there are very few correspondences, as can be seen the top part of Figure 7. While the number of correspondences is reduced even further, however, they are not eliminated completely – it would probably need longer paths (and thus many more fragments) to effect this.

In Swamidass’s work two kernels are used for classification – one based on the well-established Tanimoto-similarity [16]

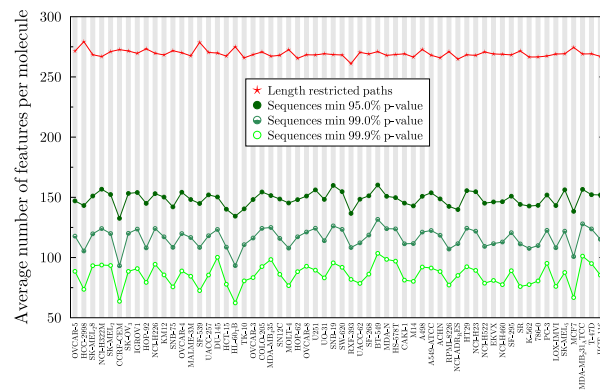


Fig. 9. Average number of fragments per molecule mined with three different significance thresholds and as length restricted paths of frequency 1.

and a so-called *Min-Max-Kernel*. While the latter gives slightly better results w.r.t. predictive accuracy, it is evaluated on a representation of the data that not only denotes absence/presence of substructures but also the number of times they occur in a molecule. Since the semantic information of significant fragments is such that only their *presence* correlates with an activity, we do not adopt this representation and thus do not compare against the Min-Max-Kernel.

The lower half of Figure 7 also lists the average AUC the SVM achieved on *gfps* using length-restriction. As we expected, using length-restricted paths with minimum support of one leads to slightly more useful feature sets but at the cost of significantly larger *fps*. In fact, while the AUC increase is not significant for most data sets (only 13/60 according to the Wilcoxon test), Figure 9 shows that the average number of fragments used to describe a single molecule effectively doubles. The gain derived from increasing the amount of features is thus affected by diminishing returns.

Those fragments are relevant in terms of the similarity of molecules yet do not capture any tendencies in the data themselves. The *kernel matrix* gives a global view of the similarity of molecules and the SVM is used to discover the underlying phenomena. Figure 10 shows that the lowest scores for fragments derived in the length-restricted approach are clearly non-significant and it would be hard to base actual understanding of the data on them. It also shows that the scores of the worst sequence and worst tree included in the top-thousand graphs are virtually indistinguishable from each other and from what is considered the worst graph. Finally, the score of the 1000th-worst tree is marginally worse than the score of the 1000th-worst graph, indicating that most of the top-1000 graphs are in fact trees.

V. CONCLUSIONS

We performed an empirical evaluation to gain insight into the reasons for the superiority of sequential molecular fragments as features for classification, compared to complex ones. We find that the reason lies in the greater diversity of sequences which leads to a more distinctive encoding of instances, effectively giving classifiers a better representation

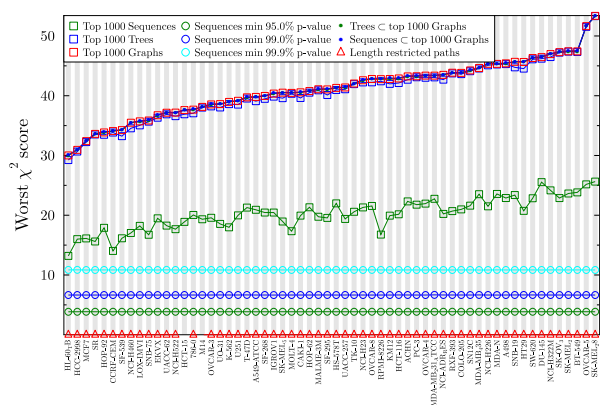


Fig. 10. Worst score for fragments mined with three different significance thresholds and as length restricted paths of frequency 1.

to work with. A straight-forward way of improving the encoding lies in increasing the number of fragments used. Our experiments show, however, that there is *always* need for a far greater number of trees/graphs than sequences. As these structures are also computationally more expensive to enumerate, this leads to vastly increased computational complexity. Our results indicate that this should be avoided.

Enumerating a subset of all sequences that cover at least one molecule produces an effective feature set but also a very large one. Also, these fragments are hard to interpret outside of their use in a pairwise similarity measure. In contrast to this, fragments that are selected based on their correlation with the target can be ranked based on their score and the most interesting ones inspected and interpreted by an end user. While it would be possible to evaluate all size restricted fragments on the data and perform a similar ranking, this will be computationally expensive due to their large number.

Our experiments also indicate a clear trade-off between the number of fragments (which can be set by adjusting the k in top- k mining or the minimum significance threshold) and the quality of the feature set. Given existing results, it is unlikely that similar clear-cut effects would appear when changing the length or minimum support of length restricted paths.

Redundancy among complex patterns could be reduced explicitly, e.g. in a post-processing step. We have suggested a technique that achieves this [17] and since the fragments can be considered features, *feature selection* techniques are applicable [18], but this would again require the mining of a *large* set of trees or graphs. It would need to be larger than a set of sequential patterns that could be used without post-processing, increasing computational complexity significantly.

Given these arguments, class-correlated sequences seem to be the best choice for the large-scale mining of molecular fragments as features for SAR prediction.

An alternative lies in iterative approaches, in which patterns are mined and data manipulated [15], [19], [20], or redundancy with already found patterns made part of the quality function [21]. Effective, very compact pattern sets can be mined in this way. So far there is however no clear understanding about whether these feature sets would be competitive with large sets

of sequences for classification. Additionally, due to sequences being graphs themselves, as explained in the introduction, it seems quite possible that such a mining operation would in the end once again give rise to a set of sequential fragments.

ACKNOWLEDGEMENTS

We would like to thank Kurt De Grave for his invaluable help in preparing this manuscript. Albrecht Zimmermann was supported by the Fonds Wetenschappelijk Onderzoek (FWO) by the time of writing.

REFERENCES

- [1] T. Washio and H. Motoda, "State of the art of graph-based data mining," *SIGKDD Explor. Newsl.*, vol. 5, pp. 59–68, July 2003.
- [2] S. J. Swamidass, J. H. Chen, J. Bruand, P. Phung, L. Ralaivola, and P. Baldi, "Kernels for small molecules and the prediction of mutagenicity, toxicity and anti-cancer activity," in *ISMB (Supplement of Bioinformatics)*, pp. 359–368, 2005.
- [3] D. Haussler, "Convolution kernels on discrete structures," TR, 1999.
- [4] T. Gärtner, P. A. Flach, and S. Wrobel, "On graph kernels: Hardness results and efficient alternatives," in *COLT* (B. Schölkopf and M. K. Warmuth, eds.), vol. 2777 of *LNCS*, pp. 129–143, Springer, 2003.
- [5] K. M. Borgwardt and H.-P. Kriegel, "Shortest-path kernels on graphs," in *ICDM* (J. Han, B. W. Wah, V. Raghavan, X. Wu, and R. Rastogi, eds.), (Houston, Texas, USA), pp. 74–81, IEEE, Nov. 2005.
- [6] N. Shervashidze and K. M. Borgwardt, "Fast subtree kernels on graphs," in *NIPS* (Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, eds.), pp. 1660–1668, Curran Associates, Inc., 2009.
- [7] S. Menchetti, F. Costa, and P. Frasconi, "Weighted decomposition kernels," in *ICML* (L. D. Raedt and S. Wrobel, eds.), vol. 119 of *ACM*, pp. 585–592, ACM, 2005.
- [8] N. Wale, I. A. Watson, and G. Karypis, "Comparison of descriptor spaces for chemical compound retrieval and classification," *Knowl. Inf. Syst.*, vol. 14, no. 3, pp. 347–375, 2008.
- [9] F. Costa and K. D. Grave, "Fast neighborhood subgraph pairwise distance kernel," in *ICML* (J. Fürnkranz and T. Joachims, eds.), pp. 255–262, Omnipress, 2010.
- [10] L. Schietgat, F. Costa, J. Ramon, and L. D. Raedt, "Effective feature construction by maximum common subgraph sampling," *Machine Learning*, vol. 83, no. 2, pp. 137–161, 2011.
- [11] G. B. McGaughey, R. P. Sheridan, C. I. Bayly, J. C. Culberson, C. Kreatsoulas, S. Lindsley, V. Maiorov, J.-F. Truchon, and W. D. Cornell, "Comparison of topological, shape, and docking methods in virtual screening," *JCIM*, vol. 47, no. 4, pp. 1504–1519, 2007.
- [12] C. Helma, ed., *Predictive Toxicology*. CRC Press, 2005.
- [13] C. Helma, T. Cramer, S. Kramer, and L. D. Raedt, "Data mining and machine learning techniques for the identification of mutagenicity inducing substructures and structure activity relationships of noncongeneric compounds," *JCIM*, vol. 44, no. 4, pp. 1402–1411, 2004.
- [14] B. Bringmann, A. Zimmermann, L. De Raedt, and S. Nijssen, "Don't be afraid of simpler patterns," in *10th PKDD* (J. Fürnkranz, T. Scheffer, and M. Spiliopoulou, eds.), pp. 55–66, Springer, 2006.
- [15] B. Bringmann and A. Zimmermann, "Tree² - Decision trees for tree structured data," in *9th PKDD* (A. Jorge, L. Torgo, P. Brazdil, R. Camacho, and J. Gama, eds.), pp. 46–58, Springer, 2005.
- [16] D. J. Rogers and T. T. Tanimoto, "A computer program for classifying plants," *Science*, vol. 21, pp. 1115–1118, oct 1960.
- [17] B. Bringmann and A. Zimmermann, "The chosen few: On identifying valuable patterns," in *ICDM* (N. Ramakrishnan and O. Zaiane, eds.), pp. 63–72, IEEE Computer Society, 2007.
- [18] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *JMLR*, vol. 3, pp. 1157–1182, 2003.
- [19] M. Thoma, H. Cheng, A. Gretton, J. Han, H.-P. Kriegel, A. J. Smola, L. Song, P. S. Yu, X. Yan, and K. M. Borgwardt, "Near-optimal supervised feature selection among frequent subgraphs," in *SDM*, pp. 1–12, SIAM, 2009.
- [20] A. Zimmermann, B. Bringmann, and U. Rückert, "Fast, effective molecular feature mining by local optimization," in *ECML/PKDD (3)* (J. L. Balcázar, F. Bonchi, A. Gionis, and M. Sebag, eds.), pp. 563–578, Springer, 2010.
- [21] U. Rückert and S. Kramer, "Optimizing feature sets for structured data," in *18th ECML* (J. N. Kok, J. Koronacki, R. L. de Mántaras, S. Matwin, D. Mladenic, and A. Skowron, eds.), pp. 716–723, Springer, 2007.